Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 12 (2014) 397 – 402

Procedia Economics and Finance

www.elsevier.com/locate/procedia

Enterprise and the Competitive Environment 2014 conference, ECE 2014, 6–7 March 2014, Brno, Czech Republic

# Deep mining of custom declarations for commercial goods

Sergei Maruev[a], Dmitry Stefanovskyi[a], Aleksey Frolov[b], Alexander Troussov[a,]*, John Curry[c]

[a]*The Russian Presidential Academy of National Economy and Public Administration*
[b]*The Federal Customs Service of Russia, ROSTEK-Pskov, Director*
[c]*Office of the Revenue Commissioners, Ireland*

**Abstract**

In our increasingly globalised world, the study of impediments to international trade is of interest to the field of international economics. This paper focuses on the particular problem of speedy and accurate processing of customs declarations. We present a novel use of graph based spreading activation algorithm for the automated processing of customer declarations for commercial goods, based on supervised learning. This method allows us to build recommender systems for use by customs officers, traders, carriers and insurers. We examine the particular use case of the recommendation to assign or not assign an armed escort to a shipping vehicle in cases of elevated risk of theft. In contrast to the usual risk based approach, this algorithm is trained solely on shipment data rather than on traditional risk indicators. This is useful as the recommendation to customs officials can be explained in terms of the make-up of a shipment and can be verified in real-time. The feasibility of the approach was tested by application to 2500 custom records collected during a continuous period of one month at eight border checkpoints between Russian Federation and two EU countries. The algorithm achieved 100 % accuracy under experimental conditions.


Selection and/or peer-review under responsibility of the Organizing Committee of ECE 2014

*Keywords:* Custom declarations; data mining; graph-based methods; spreading activation

* Corresponding author. Tel.: +353-85-8137022.
*E-mail address:* troussov@gmail.com


Selection and/or peer-review under responsibility of the Organizing Committee of ECE 2014
doi:10.1016/S2212-5671(14)00360-8

## 1. Introduction

International trade is one of the most important drivers of the global economy. Therefore, the study of impediments to this trade is of interest to the field of international economics. International trade is typically more costly than domestic trade due to the imposition of extra direct and indirect costs including tariffs, time costs due to border delays and processing costs that are exacerbated by differences in language, legal system and culture, see, for instance, Zvetkov et al. (2013) in Russian.

This paper focuses on the particular problem of speedy and accurate processing of custom declarations for the minimization of time delays and processing costs to traders and for cost minimization and improved fraud risk management for customs agencies. We present a novel algorithm for the automated processing of customer declarations for commercial goods, based on supervised learning. This algorithm allows us to build recommender systems for use by customs officers, traders, carriers and insurers. The system outputs a recommended action given inputs related to the trade event. Example recommended actions could be the assignment of an armed escort to a shipping vehicle in cases of elevated risk of theft or the raising of a red flag in cases of potential fraud or error in a declaration. The system is designed to integrate into the human customs work-flow and, rather than being a black box, provides reasons for each recommendation to the end user.

We investigate the feasibility of the approach by its application to 2500 custom records collected during a continuous period of one month at eight border checkpoints between Russian Federation and two EU countries.

The rest of the paper is organized as follows. In section 2 we provide an overview of raw data. In section 3 we describe generic method of network data representation and algorithms of mining. We also describe the task of recommending armed escort and details of the application of our generic method to this particular task. In section 4 we present the results of experimental validation of for the recommending armed escort. Finally, section 5 describes the conclusions and future work.

## 2. Raw data description

The raw data for this study originates with traders shipping goods into the Russian Federation through land borders. They comprise a description of the itemized contents of a shipment of goods in a particular vehicle (always a truck in this study). These data are used to produce custom goods declarations and to compute taxes. When the truck crosses the border, the data become part of the custom service's electronic data archive.

Each item record describes a specific type of goods, and has the following numeric and alphanumeric fields:

- № – the number of item record;
- Document_id (DI) – an unique identification number of the document for one shipment;
- Escort (E) – the value 1 means that the item is located in a truck which should be accompanied by armed escort;
- Consignee_id (C1), Consignor_id (C2), Carrier_id (C3) – identification numbers of consignee, consignor, and carrier;
- GoodsTNVEDCode – ten digits code for the commodity. This nomenclature is used in the Customs Union of Belarus, Kazakhstan, and Russia† and is also consistent with the codes used in the European Union‡
- GoodsDescription – goods description;
- GrossWeight (GW) –gross weight in kgs;
- InvoicedCost – invoiced cost;
- CurrencyCode (CC) – currency code (USD, EUR, etc);
- CurrencyRate (CR) – currency rate.

Table1 shows an example of an item record which uses Russian language description of goods related to printing machinery.

† http://en.wikipedia.org/wiki/Customs_Union_of_Belarus,_Kazakhstan,_and_Russia
‡ http://www.rusimpex.ru/Content_e/Reference/Tnv/tnv_eng.htm

Table 1. An example of goods item with the invoiced cost of 531495.03 USD.

| № | DI | E | C1 | C2 | C3 | Goods-TNVEDCode | GoodsDescription | GW | InvoicedCost | CC | CR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 385 | 389 | 785 | 8443999009 | ЧАСТИ И ПРИНАД-ЛЕЖНОСТИ ПЕЧАТНЫХ МАШИН | 7482.320 | 531495.03 | USD | 33.2474 |

All the items contained in one vehicle must have the same unique Document_id. The sets of item records with the same Document_id forms a document needed to cross the border. Table 2 shows an example of such a document with the Document_id number 11 which has toys (in Russian – ИГРУШКИ ), whisky (ВИСКИ) and stationary items notepads (БЛОКНОТЫ ДЛЯ ЗАПИСЕЙ).

Table 2. An example of the itemised document to complete custom declaration for goods transported in one vehicle. Note that all the items must have the same carrier and currency rates.

| № | DI | E | C1 | C2 | C3 | Goods-TNVED-Code | GoodsDescription | GW | InvoicedCost | CC | CR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 11 | 0 | 967 | 218 | 204 | 9503003500 | ИГРУШКИ | 159.700 | 3985.64 | EUR | 44.0129 |
| 8 | 11 | 0 | 967 | 218 | 204 | 2208308200 | ВИСКИ | 295.950 | 943.20 | EUR | 44.0129 |
| 9 | 11 | 0 | 967 | 218 | 204 | 4820103000 | БЛОКНОТЫ ДЛЯ ЗАПИСЕЙ | 15.140 | 128.64 | EUR | 44.0129 |

## 3. Data representation and mining

We represent these data as multidimensional networks, see Troussov et al. (2011). Nodes and links are typed. Nodes represent complete data fields or, in the case of the “GoodsDescription”, particular words from that field. An example of such a network is given in Figure 1 below. This network and the methods of its construction will be described in Section 3.2 Network Construction.

To demonstrate the use of our method we choose a particular task – to automatically detect whether or not the vehicle carrying the shipment needs an armed escort. We consider this task as a task of supervised machine learning. We split the data set into two halves, learn the rules based on the first half, and apply the rules to the second half (not used in the procedure of learning). We then validate the results against the known outcomes to assess the predictive power of our rules.

In our approach, learning is done in two distinctive stages. First of all, we build the network from data. Secondly, we use graph algorithms to find patterns in the network.

In this section we focus on the network construction as a means of data representation. There are different ways to represent the same data as a network with different levels of completeness and granularity. We need to preserve enough information to find particular patterns we are interested in, avoiding “noisy” data features. For instance, in certain cases the armed escort could be assigned by the border control based not on the transported goods, but because the carrier is in so called “black list” of carriers suspected in fraudulent activities. In one is interesting in automated inferring of black-listed carriers, carriers must be represented in the network, but such task hardly could be considered as a task for deep mining of custom declarations. To demonstrate the potential of our methods we choose a much more difficult task – to infer the need for armed convoy based solely on good descriptions in natural language and their GoodsTNVEDCode (the information that could be easily verified by custom officer). Fig.1 shows the fragment of the network in that reduced feature space constructed in our validation experiment.

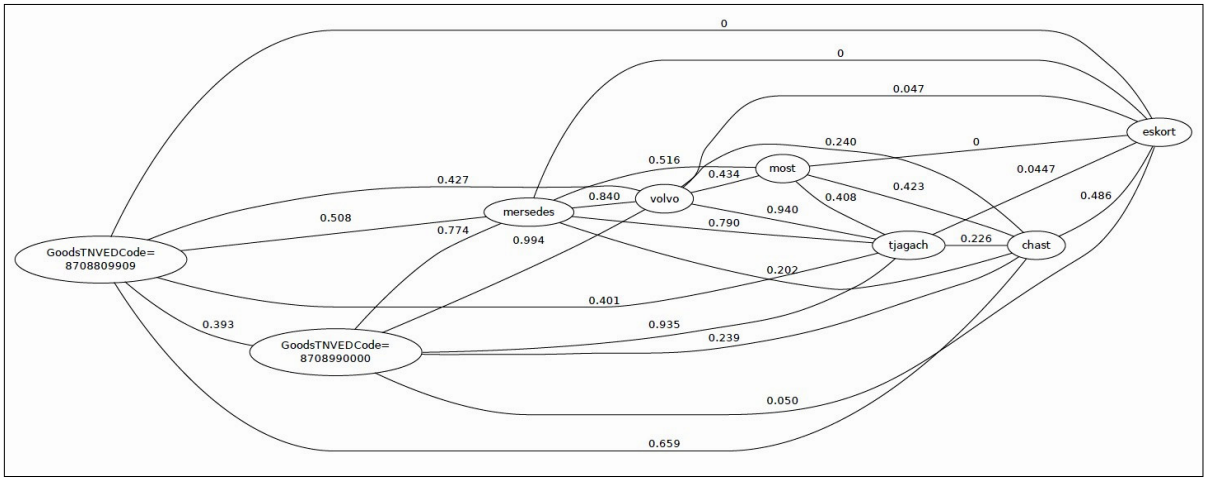


Fig. 1. The fragment of the network which represents the data from custom declarations. This network shows, for example, that the word "tjagach" ("тягач" in Russian; rooter, tractor in English) was met in good descriptions which require armed escort in 0.0447 % cases. The shipments which has in the description the word "mersedes" ("мерседес" in Russian; Mercedes-Benz international) never required armed escort, while "volvo" required escort in 0,047 % of cases. Shipments with parts, details of something ("chast", "часть" in Russian), frequently were escorted.

The risk of theft is not purely a function of the shipment value. For instance, while some truck parts might be very valuable, they are stamped with serial numbers and would be difficult to use when stolen.

### *3.1. Network construction*

The construction and the use of this network consist of the following generic steps which could be adjusted for use in other scenarios.

1. The first 2000 items description lines were used to construct the network
2. We removed all the fields except the commodity codes ('GoodsTNVEDCode"), the goods description ("GoodsDescription"), and the escort flag ("Escort").
   Their removal is not a necessary step, but we have done it in order to reduce the volume of learning data and test our algorithm under harsh conditions.
3. Each word was reduced to its normalized form by the procedure known as stemming, see, for instance, Jurafsky and Martin (2009). Stemming is an empirical natural language processing procedure allowing to map inflected and derived words into one index form; for instance, to map "fishing", "fished", and "fisher" to the form "fish". For this procedure we used Porter stemmer for Russian.
4. All the items belonging to one shipment document are merged, that is: all commodity codes; all stems from goods description and all of the escort flags (which are the same for all items since it is an attribute of shipment, not of particular items in shipment);
5. As the result, we obtained 4352 entities, which we merged into one network with 4352 nodes.
   If the two entities are met at least once in one same shipment document, the corresponding pair of nodes is connected by an arc. The weight of that arc represents how frequently the two entities corresponding to the pair of nodes are met in a shipment document, i.e. the number of co-occurrences divided by the number of documents.

### *3.2. Modeling new documents as sets of nodes on the network*

The network represents the learning data. Each new document could now be modeled as a set of nodes on this network. I.e. for each new document we need to perform steps 1-4 described above. Each new entity is mapped into a corresponding node on the network. If such a node is not found, the entity is ignored. It could not be usefully present in the model because our "learning" has no knowledge about such entities.

### *3.3. Mining*

Mining now can be done by various graph-based methods. In our experiments we used the set of operations based on the Spreading Activation Method, described in Troussov et al. (2009), and its generalization in the paper Troussov et al. (2011). For the purposes of the experiment in this paper, it works as described below.

The Spreading Activation Method has its origin in neurophysiology: "In neurophysiology interactions between neurons is modeled by way of activation which propagates from one neuron to another via connections called synapses to transmit information using chemical signals. The first spreading activation models were used in cognitive psychology to model this processes of memory retrieval." – Troussov et al. (2009). Later this framework was exploited in Artificial Intelligence as a method for searching associative, neural or semantic networks; see, for example, Crestani (1997), Aleman-Meza et al. (2003), Rocha et al. (2004).

In this paper we treat spreading activation as a scheme for iterative computations of a function on network nodes in line with papers Troussov et al. (2011) and Troussov et al. (2011a). On each iteration the value of $F_{n+1}(v)$ is computed depending on the values of the function $F_n$ on the nodes connected to the node $v$.

Since we want to recommend the assignment or non-assignment of an armed convoy for the shipment, when building the model of a new document, we remove the *Escort* flag from the model (that is different to step 2 described above, where the escort flag has been left in). Now our task is to compare a model of a new document, which we may call *Model 2* to the model of existing documents, which we may call *Model 1*. *Model 1* contains the Escort flag nodes *"Yes"* and *"No"*. In other words, we now need to evaluate the relations between *Model 1* and *Model 2*, i.e. to compute the cumulative strengths of connections between the model of a new document and the single node Escort flag *"Yes"*. In terms of the use of spreading activation, we activate *Model 1* (that is we put the initial activation at all network nodes comprising the *Model 1*) and compute how much activation comes to the nodes in *Model 2*.

Depending on the task and the structural properties of the network, a few up to several dozen iterations of spreading activation are normally sufficient to achieve the goal. We found that one iteration of spreading activation is enough for our purposes. In other words, the number of arcs and their weights between the model and the node *Escort* is a good predictor that the shipment requires an armed escort. The larger the weights, likelier it is that the shipment requires the convoy. The effect of the number of arcs here is much less evident, because, logically, a high number of arcs with small weights indicate that the shipment doesn't require a convoy.

To aggregate weights one can use the arithmetic mean of the weights of arcs. However, as we found in our experiment, arcs with high weights close to *1.0* were important, while arcs with small weights were not reliable predictors. Therefore, instead of arithmetic mean for $n$ real numbers $x_1, x_2, \ldots, x_n$ representing weights, we used the mean computed as the $L^p$ –norm of the vector $\{x_1, x_2, \ldots, x_n\}$ with the parameter $p$ empirically taken with the value *3.5* to favor links with high weights and to ignore links with very small weights.

$$||\mathrm{x}||_{\mathrm{p}} = (|x_1|^{\mathrm{p}} + |x_2|^{\mathrm{p}} + \ldots + |x_n|^{\mathrm{p}})^{1/p} \quad (1)$$

## 4. Evaluation

Speaking in terms of machine learning, we used half of the data for learning, half for the evaluation. In our approach, learning is done in two stages. First of all, we build the network from data, and there are different ways to represent the data with different levels of completeness and granularity. Secondly, we used graph algorithms to find patterns in the network.

Each new document is mapped onto the network, and the strength of its connection to the escort flag was computed and compared with the value of the escort flag in the document. The results were robust, and the choice of the threshold was not a problem. 100 % of the flags were recognized correctly.

## 5. Conclusions and future work

Our paper argued for the use of a network form of data representation and spreading activation based algorithms for mining custom goods declaration. We introduced the method for modeling the data as a multidimensional network, where nodes represent various codes and alphanumeric fields, as well as the terms extracted from the goods description in natural language. We briefly outlined various scenarios of finding patterns in data and their use to build recommender systems for use by customs officers, traders, carriers and insurers. The system outputs a recommended action given inputs related to the trade event. Example recommended actions could be the assignment of an armed escort to a shipping vehicle in cases of elevated risk of theft or the raising of a red flag in cases of potential fraud or error in a declaration. The system is designed to integrate into the human customs work-flow and, rather than being a black box, provides reasons for each recommendation to the end user.

We validated our approach on a particular task of finding patterns by application of our technique to 2500 custom records collected during a continuous period of one month at eight border checkpoints between Russian Federation and two EU countries. Representing the data for the first 1250 records in the network form, and applying spreading activation-based algorithm (according to Troussov et al. (2009) spreading activation might be considered as a method for soft clustering on networks), we were able to find pattern which describe if a shipping vehicle requires assignment of an armed escort. We applied this pattern to the rest of the data, which were not used in the procedure of finding pattern, and in 100 percent cases predictions given by that pattern were accurate.

In our experiments we used only the data from custom declaration. We are currently running experiments to add the external relevant knowledge, such as the hierarchical structure of the commodity nomenclature and the model of misspellings in good description. The network form of representation provides ease of merge of heterogeneous information; any such external knowledge could be added on the top of the network obtained from data as new nodes and new weighted arcs.

## References

Aleman-Meza, B., Halaschek, C., Arpinar, I., Sheth, A., 2003. Context-Aware Semantic Association Ranking. Proceedings of SWDB'03, Berlin, Germany, 33–50.

Crestani, F., 1997. Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review, 11(6), 453–482.

Jurafsky, D. and Martin, J.H., 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (2ed.), Prentice Hall.

Rocha, C, Schwabe, D., Poggi de Aragao, M., 2004. A Hybrid Approach for Searching in the Semantic Web. Proceedings of the 13th international conference on World Wide Web, May 17–20, 2004, New York, NY, USA, 374–383.

Troussov, A., Darena, F., Zizka, J., Parra, D., and Brusilovsky, P., 2011a. "Vectorised Spreading Activation Algorithm for Centrality Measurement". Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis. sv. LIX, no. 7, s. 469–476. ISSN 1211-8516.

Troussov, A., Jundge, J., Alexandrov, M., and Levner, E., 2011. Social Context as Machine-Processable Knowledge. Proceedings of the International Conference on Intelligent Information and Engineering Systems INFOS 2011, Rzeszów - Polańczyk, Poland, pp. 104–114, ISBN: 978-954-16-0053-5.

Troussov, A., Levner, E., Bogdan, C., Judge, J., Botvich, D., 2009. Spread of Activation Methods. In Dynamic and Advanced Data Mining for Progressing Technological Development, Y. Xiang and S. Ali (eds) IGI Global.

Zvetkov, V, Zoidov, K, and Medkov, A., 2013. In Russian. Цветков, В, Зоидов, К., Медков, А. О возможности и целесообразности организации транзита через Россию грузов между странами Тихоокеанского региона и Европы. Депонирована в системе Соционет, февраль 2013 г. Retrieved February 27, 2014, from http://www.cemi.rssi.ru/mei/articles/tsvetkov-and13-01.pdf